AD-A033 442 PURDUE UNIV LAFAYETTE IND SCHOOL OF ELECTRICAL ENGI--ETC F/G 12/2
A BAYESIAN COMPARISON OF DIFFERENT CLASSES OF DYNAMIC MODELS US--ETC(U)
SEP 76 R L KASHYAP
UNCLASSIFIED TR-EE76-40 AFOSR-TR-76-1244 AF-AFOSR2661-74 NL

| OF |
AD A033442

END

DATE FILMED
2-77

A BAYESIAN COMPARISON OF DIFFERENT CLASSES OF DYNAMIC MODELS USING EMPIRICAL DATA

R. L. Kashyap

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

TR-EE 76-40
September 1976

This work was supported by Air Force Office of Scientific Research
under Grant 74-2661.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR-76-1244
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): A BAYESIAN COMPARISON OF DIFFERENT CLASSES OF DYNAMIC MODELS USING EMPIRICAL DATA.

5. TYPE OF REPORT & PERIOD COVERED: Interim rept.

6. PERFORMING ORG. REPORT NUMBER: TR-EE 76-48

7. AUTHOR(s): R. L. Kashyap

8. CONTRACT OR GRANT NUMBER(s): AF-AFOSR-2661-74

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Purdue University
School of Electrical Engineering
West Lafayette, IN 47907

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A2

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Office of Scientific Research/NM
Bolling AFB, Washington, DC 20332

12. REPORT DATE: September 1976

13. NUMBER OF PAGES: 35

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 39p.

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper deals with the Bayesian methods of comparing different types of dynamical structures for representing the given set of observations. Specifically, given that a given process y(t) obeys one of r distinct stochastic difference equations each involving a vector of unknown parameters, we compute the posterior probability that a set of observations {y(1),...,y(N)} obey the i-th equation, after making suitable assumptions about the prior probability distribution of the parameters in each class. The difference equations can be nonlinear in the variable y but should be linear in the parameter vector

next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 292 000

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20 Abstract

cont'd

it. Once the posterior probability is known, we can find a decision rule to choose between the various structures so as to minimize the average value of a loss function. The optimum decision rule is asymptotically consistent and gives a quantitative explanation for the 'principle of parsimony' often used in the construction of models from empirical data. The decision rule answers a wide variety of questions such as the advisability of a nonlinear transformation of data, the limitations of a model which yields a perfect fit to the data (i.e., zero residual variance) etc. The method can be used not only to compare different types of structures but also to determine a reliable estimate of spectral density of process. We compare the method in detail with the hypothesis testing method, maximum entropy spectral analysis method and other methods and give a number of illustrative examples.

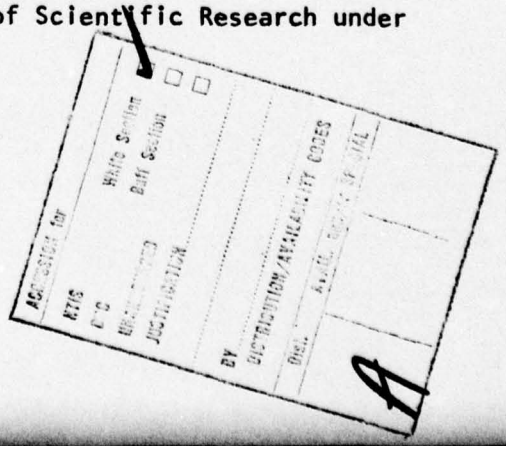
A BAYESIAN COMPARISON OF DIFFERENT CLASSES OF DYNAMIC MODELS USING EMPIRICAL DATA*

R. L. Kashyap
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

TR-EE 76-40

September 1976

*This work was supported by Air Force Office of Scientific Research under Grant 74-2661.



A Bayesian Comparison of Different Classes of Dynamic Models Using Empirical Data*

R. L. Kashyap†

Abstract

This paper deals with the Bayesian methods of comparing different types of dynamical structures for representing the given set of observations. Specifically, given that a given process $y(\cdot)$ obeys one of r distinct stochastic difference equations each involving a vector of unknown parameters, we compute the posterior probability that a set of observations $\{y(1),\ldots,y(N)\}$ obey the ith equation, after making suitable assumptions about the prior probability distribution of the parameters in each class. The difference equations can be nonlinear in the variable y but should be linear in the parameter vector in it. Once the posterior probability is known, we can find a decision rule to choose between the various structures so as to minimize the average value of a loss function. The optimum decision rule is asymptotically consistent and gives a quantitative explanation for the 'principle of parsimony' often used in the construction of models from empirical data. The decision rule answers a wide variety of questions such as the advisability of a nonlinear transformation of data, the limitations of a model which yields a perfect fit to the data (i.e., zero residual variance) etc. The method can be used not only to compare different types of structures but also to determine a reliable estimate of spectral density of process. We compare the method in detail with the hypothesis testing method, maximum entropy spectral analysis method and other methods and give a number of illustrative examples.

---

*This work was supported by Air Force Office of Scientific Research under Grant 74-2661.

†School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907.

I. INTRODUCTION

For determining an appropriate representation for a given time series, one usually assumes a certain structure involving a set of parameters and the values of these parameters are estimated from the given data. For many of the geophysical or hydrological time series, the physics of the problem is not well understood to specify a unique structure for the stochastic process. In these cases, one of the main reasons for the construction of models is for understanding the dynamics of the process. Hence instead of restricting ourselves to one particular structure like autoregressive structure, a linear combination of orthogonal polynomials in time t or a fourier series, we should endeavor to quantitatively compare the validity of these widely different structures for representing the given data. For every structure, we should try to determine the probability that the given set of measurements could have come from that structure and choose that structure which has the highest probability.

We will first clarify the notion of structure and model because these words are used in widely differing contexts. Consider a stochastic process $y(\cdot) \in \mathcal{Y}$ which obeys a stochastic difference equation (1.1).

$$f_i(y(t)) = g_i(t, y(t-1), \ldots, y(t-m_i), \underset{\sim}{\theta}) + w(t), \qquad (1.1)$$

where f_i etc. obey the following assumptions (A1) - (A3).

(A1) $f_i(y)$ is any differentiable function of y such as $\ln y + C$, $y + C$, $y^2 + C$, etc.; $f_i: y \to R$.

The constant C is chosen so that the empirical mean of $f_i(y)$ with the given set of observations $\xi(N) = (y(1), \ldots, y(N))$ is zero, i.e., $C = (1/N) \sum_{t=1}^{N} f_i(y(t))$.

(A2) Let $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_{n_i})^T \in R^{n_i}$. g_i is any function linear in $\underset{\sim}{\theta}$, but not necessarily linear in y(t-1), t etc. m_i is any integer greater than or equal to zero.

$g_i: \tau \times \mathcal{Y}^{m_i} \times R^{n_i} \to R$.

$\tau = \{1,2,3,\ldots\}$.

when $m_i = 0$, the function g_i is defined as a map g_i: $\tau \times R^{n_i} \to R$.

The parameters m_i and n_i characterizing the function g_i will be of particular interest later in developing decision rules.

(A3) $\{w(\cdot)\}$ is a sequence of I.I.D. variables with distribution $N(0,\rho)$ and $w(t)$ is independent of $y(t-j)$, $j \geq 1$, $0 \leq \rho < \infty$.

A model is a 4 tuple $\{f_i, g_i, \underset{\sim}{\theta}, \rho\}$ uniquely associated with the difference eq. (1.1). A process $y(\cdot)$ is said to obey a model $\{f_i, g_i, \underset{\sim}{\theta}, \rho\}$ if it obeys the associated difference equation (1.1).

Let C_i be a class of models having the same functions f and g but differing in $\underset{\sim}{\theta}$ or ρ.

$$C_i \triangleq \{f_i, g_i, m_i, n_i, \ \}$$
$$= \{(f_i, g_i, \underset{\sim}{\theta}, \rho): \ \underset{\sim}{\theta} \in \Theta_i, \ 0 \leq \rho < \infty\}, \tag{1.2}$$

where

$$\Theta_i = \{\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_{n_i})^T, \ \theta_i \neq 0 \ \ \forall_i, \ \underset{\sim}{\theta} \in R^{n_i}\} \tag{1.3}$$

C_i is labelled as the ith class or ith structure. Two classes C_i and C_j are said to be mutually exclusive if they differ in the functions f or g, in a nontrivial way i.e., there is no process which obeys a model in C_i as well as a model in C_j.

Given r mutually exclusive classes or structures $C_1, \ldots, C_r$ and a set of observations $\xi = \{y(1), \ldots, y(N)\}$ our intention is to develop a decision rule to assign the observation set to one of the classes among C_i, $i = 1, \ldots, r$ so as to minimize a suitable criterion function and also determine the probability of error associated with the decision. The loss function can be chosen to reflect the particular needs of the problem, such as forecasting or estimation of spectral density.

We recall that the only restriction on the classes C_i is that the functions g_i, $i = 1,2,\ldots$ must be linear in $\underset{\sim}{\theta}$. Hence the theory allows us to simultaneously compare widely different structures or classes such as (i) the class of autoregressive (AR) models of order m_1 in y, (ii) the class of AR models of order m_2 in ln y, (iii) the class of models made of m_3 orthogonal polynomials in the variable t, (iv) class of harmonic models involving certain frequencies, using the same data.

Problems of this type are referred to as compound hypothesis testing in the statistical literature [2] and there is no general theory to handle them. There are specific tests for comparing specific pairs of classes such as classes of autoregressive models of 2 different orders. Even here the solution is unsatisfactory for a number of reasons such as (i) the choice of arbitrary quantities like significance levels (ii) lack of any measure of the type II error involved, (the chosen significance level places an upperbound on the probability of type I error) (iii) the intransitivity of the decision rule and its lack of any optimality properties etc. There are also many other adhoc rules such as likelihood ratio rule [2] or MAIAC [3] which do not usually possess asymptotic consistency or give a measure of the probability of error of the decision.

Some common problems occurring in the analysis of empirical time series are really problems regarding the appropriateness of the different types of structure for explaining the data. It is being increasingly realized that the traditional methods of spectral analysis involving the use of window functions like those of Bartlett or Hamming [6] often lead to misleading inferences. To overcome such objections, geophysicists use a special method of spectral analysis known as maximum entropy spectral analysis (MESA) whose inferences have greater reliability than those of the traditional methods. The method effectively assumes an autoregressive (AR) structure for the process. If we pose the problem as one of comparing different classes of models having different orders of AR models as done here,

we will obtain a solution which overcomes many of the disadvantages of the MESA method, without sacrificing the advantages. Moreover, we can choose the decision rule to minimize a loss function which explicitly reflects the needs in the context namely minimizing the errors in the estimation of spectral density of the process. Similar problems are faced in the spectral analysis as applied to the processing of acoustic signals also [7].

In our paper we will first compute $P(C_i|\xi)$, the posterior probability of the given data having been generated by some model in C_i, for every $i = 1,\ldots,r$, assuming a suitable prior distribution for the classes and the various parameters. Subsequently we will derive optimal decision rules according to various types of criteria and discuss the asymptotic consistency of the decision rule. We will give many types of examples such as (i) whether a nonlinear transformation of the data will yield a better representation of the data, (ii) when is a polynomial fit of the data (or a fourier series representation of the data) with zero residual variance inferior to a relatively small order autoregressive representation etc., and point out the superiority of the method to existing methods of comparison like hypothesis testing [1,2], maximum entropy spectral analysis [4,5,10], etc.

Even if all the classes have the same prior probability, the posterior probability of class C_i being the correct class for the observation set $\xi(N)$ having N observations is, under appropriate assumptions

$$P(C_i|\xi(N)) = \frac{\exp[0.5\, h_i(\xi(N))]}{\sum_{k=1}^{r} \exp[0.5\, h_k(\xi(N))]}, \quad i = 1,\ldots,r \tag{1.4}$$

$$h_i(\xi(N)) = 2N\,\overline{\ln f_i'} - (N-m_i)\ln(\rho_i + (\rho_{fi}-\rho_i)/N) - n_i \ln N - m_i \ln \rho_{fi} + m_i - (1/\rho_{fi}) \sum_{t=1}^{m_i} (f_i(t))^2 + O(1/N) \tag{1.5}$$

$$i = 1,2,\ldots,r$$

where $\overline{\ln f_i'}$ is the empirical mean of the function $\ln[d f_i(y)/dy]|y = y(t)$, ρ_{fi} is the empirical variance of $f_i(y(t))$ and ρ_i is the residual variance. The expression (1.5) clearly brings out the adverse effects of dealing with classes with large n_i, i.e., having too many parameters to be estimated. The role played by the lag variable m_i and the number of estimated parameters n_i is quite different in general even though in the widely discussed case of autoregressive processes $m_i = n_i$. Further, even if $\rho_i = 0$, as in polynomial models, $P(C_i|\xi(N))$ is still finite. These and other features will be discussed extensively in the subsequent sections.

II. ASSUMPTIONS

For simplicity we will denote a class C_i by the 4 tuple $[f_i, g_i, m_i, n_i]$. The suppressed term Θ_i will be mentioned wherever necessary. Further $f_i(t) \triangleq f_i(y(t))$. Similarly the function g_i will be written as $g_i(t,\underset{\sim}{\theta})$ suppressing the other arguments. Moreover g_i can be written as:

$g_i(t,\underset{\sim}{\theta}) = (\underset{\sim}{z}_i(t-1))^T\underset{\sim}{\theta}$, where $\underset{\sim}{z}_i$ is independent of $\underset{\sim}{\theta}$,

$\underset{\sim}{z}_i^T(t-1) = \underset{\sim}{\nabla}_{\underset{\sim}{\theta}} g_i(t,\underset{\sim}{\theta})$, n_i - vector

We consider r classes C_i, $i = 1,\ldots,r, C_i = [f_i, g_i, m_i, n_i]$ where f_i and g_i obey the conditions (A1) - (A3). We will make the following additional assumptions on the functions f, g etc.

(A4) The functions f_i and f_j and g_i and g_j obey at least one of the following conditions for every possible pair (i,j), $i \neq j$, $i,j = 1,\ldots,r$

(i) $f_i(y) \neq kf_j(y) + C$ for any k,C and almost all y

(ii) For every $\underset{\sim}{\theta} \in \Theta_i$, there does not exist a $\underset{\sim}{\theta}' \in \Theta_j$ satisfying the following relation

$$g_i(t,y(t-1),\ldots,y(t-m_i),\underset{\sim}{\theta}) = g_j(t,y(t-1),\ldots,y(t-m_j),\underset{\sim}{\theta}')$$

(A5) If a process $y(\cdot)$ belongs to one of the classes C_i, $i = 1,2,\ldots,r$, then $[1/N \sum_{t=m_i+1}^{N} \underset{\sim}{z}_i(t-1)\underset{\sim}{z}_i^T(t-1)]$ is finite and positive definite for all N and $i = 1,\ldots,r$.

We need the following assumptions regarding the prior probability distribution of $\underset{\sim}{\theta}$ and ρ and the probability distribution of the initial conditions for eq. (1.1).

(A6) Let $x(t) = f_i(y(t))$. The m_i initial values $y(1),\ldots,y(m_i)$ of a process $y(\cdot)$ obeying any model in C_i obeys the following normal density

$p(x(1),\ldots,x(m_i)) \sim N[\underset{\sim}{0},\underset{\sim}{R}_i]$

where $\underset{\sim}{0}$ is the null vector of dimension m_i and $\underset{\sim}{R}_i$ is a $m_i x m_i$ covariance matrix.

(A7) If $y(\cdot)$ obeys a model in C_i involving m_i lagged variables, then the prior density of $\underset{\sim}{\theta}$ and ρ obeys the following relation:

$$p(\underset{\sim}{\theta},\rho|y(1),\ldots,y(m_i),C_i) = p(\underset{\sim}{\theta},\rho|C_i)$$

(A8) The variable ρ has the following probability density: $p(\rho|C_i) = \alpha/\rho$, $\alpha > 0$, valid for all i.

(A9) The variable $\underset{\sim}{\theta} \in (H)_i$ has the following conditional prior density given ρ:

$$p(\underset{\sim}{\theta}|\rho,C_i) \sim N(\underset{\sim}{\theta}_{0i},\underset{\sim}{S}_{0i}\rho)$$

where $\underset{\sim}{S}_{0i}$ is a

(A10) The prior probability of class C_i is $P(C_i)$, $i = 1,2,\ldots,r$, $0 < P(C_i) < 1$, $\Sigma_{i=1}^{r} P(C_i) = 1$.

We will discuss the assumptions. The assumption (A4) and condition (1.3) are sufficient for the classes $C_1,\ldots,C_r$ to be mutually exclusive, i.e., there does not exist a process $y(\cdot)$ which obeys 2 different models in 2 different classes. (A5) is needed for the existence of the matrices occurring in the optimal decision function. To understand (A6) and (A7), we should note that we have not made any assumption about the stationarity of the process y obeying any model in C_i for any i. Specifically, if $y(\cdot)$ belongs to C_i, then $\{y(1),\ldots,y(N)\}$,

the observation set is generated as follows: $y(1),\ldots,y(m_i)$ act as initial conditions. Based on $y(1),\ldots,y(m_i)$ and $w(m_i+1)$, $y(m_i+1)$ is generated from (1.1), and $y(t)$, $t > m_i+1$ are generated recursively using (1.1). Clearly the initial conditions $y(1),\ldots,y(m_i)$ cannot throw any light on the parameters $\underset{\sim}{\theta}$ and ρ which characterize that particular model in C_i obeyed by $y(\cdot)$. This statement is the assumption (A7).

The assumption (A3) yields the conditional probability density $p(y(m_i+1),\ldots, y(N)|y(1),\ldots,y(m_i)\underset{\sim}{\theta},\rho,C_i)$ as shown below. This expression in conjunction with (A6) and (A7) yields the joint probability density of all the observations namely $p(y(1),\ldots,y(N)|\underset{\sim}{\theta},\rho,C)$. The details are indicated below.

$$p(x_i(m_i+1),\ldots,x_i(N)|x_i(1),\ldots,x_i(m_i),\underset{\sim}{\theta},\rho,C_i)$$

$$= \prod_{t=m_i+1}^{N} p(x_i(t)|x_i(t-1),\ldots,x_i(1),\theta,\rho,C), \text{ by (A3)}$$

$$= \prod_{t=m_i+1}^{N} \frac{1}{\sqrt{2\pi\rho}} \exp[-1/2\rho(x_i(t) - g_i(t,\theta))^2], \text{ by (A3)} \qquad (2.1)$$

Transforming the variables x into y by the relation $x_i(t) = f_i(y(t))$, (2.1) yields:

$$p(y(m_i+1),\ldots,y(N)|y(1),\ldots,y(m_i),\underset{\sim}{\theta},\rho,C_i)$$

$$= \prod_{t=m_i+1}^{N} (f_i'(t)/\sqrt{2\pi\rho})\exp[-1/2\rho(f_i(y(t)) - g_i(t,\underset{\sim}{\theta}))^2], \qquad (2.2)$$

where $f_i'(t) = (d\, f_i(y)/dy)|y = y(t)$

$$p(y(1),\ldots,y(m_i)|\theta,\rho,C_i) = p(y(1),\ldots,y(m_i)|C_i), \text{ by (A7)},$$

$$= \frac{\prod_{t=1}^{m_i} f_i'(t)}{(2\pi)^{m_i/2}\,(\det R_i)^{1/2}} \exp[-(1/2)||(f_i(1),\ldots,f_i(m_i)||^2_{R_i^{-1}}], \text{ by (A6)} \qquad (2.3)$$

Multiplying the expressions (2.2) and (2.3) yields the required density $p(y(1),\ldots,y(N)|\underset{\sim}{\theta},\rho,C_i)$.

The prior density of $\underset{\sim}{\theta}$ given ρ in (A9) is so chosen that the posterior density $p(\theta|\rho,\xi,C_i)$ is also Gaussian. There is an extensive literature on the choice of the prior density of ρ. The density $p(\rho)$ given in (A8) is commonly used in statistical literature and can be defended on many different grounds [8]. However it is called improper since $\int_{a_1}^{\infty} p(\rho)d\rho$ does not exist for $a_1 = 0$. But in computing the posterior density, we can allow the limit a_1 to be zero.

We will offer some suggestions for the choice of $\underset{\sim}{\theta}_{0i}$ and $\underset{\sim}{S}_{0i}$, occurring in the prior density $p(\underset{\sim}{\theta}|\rho,C_i)$. The only guideline available for the choice of $\underset{\sim}{S}_{0i}$ is that it be relatively large in view of the great initial uncertainty regarding the value of $\underset{\sim}{\theta}$ appropriate for the given data. Even though the effect of $\underset{\sim}{\theta}_{0i}$ and $\underset{\sim}{S}_{0i}$ are asymptotically insignificant on the optimal decision, still the arbitrary choice of θ_{0i} and S_{0i} for various i may make the computation of the $p(C_i|\xi)$ more cumbersome than it need be. Accordingly we propose that the following choice for S_{0i}, i = 1,...,r.

$$\text{(A11)} \qquad \underset{\sim}{S}_{0i} = [1/N-m_i \sum_{t=m_i+1}^{N} \underset{\sim}{z}_i(t-1)\underset{\sim}{z}_i^T(t-1)]^{-1}$$

The choice in (A11) is unconventional in the Bayesian literature since it depends on the observations. However, we will show later that such a choice implies the following expression for the posterior variance of $\underset{\sim}{\theta}$ given ξ and ρ

$$\text{Var}[\theta|\xi,\rho,C_i] \triangleq \underset{\sim}{S}_i = \underset{\sim}{S}_{0i}/(N-m_i+1) \qquad (2.4)$$

our choice is reasonable since $\underset{\sim}{S}_{0i}\rho$ is much larger than the posterior variance $S_i\rho$. Any other choice for the $\underset{\sim}{S}_{0i}$ would have made the expression for S_i more complicated than the one in (2.4).

Next let us turn our attention to the choice of the vector $\underset{\sim}{\theta}_{i0}$ occurring in $p(\theta|\rho,C_i)$. The obvious choice for $\underset{\sim}{\theta}_{i0}$ is the null vector, stated in (A12).

$$\text{(A12)} \qquad \underset{\sim}{\theta}_{i0} = [0,\ldots,0]^T, \ (n_i\text{-vector})$$

It is important to realize that (A12) is valid only if $f_i(\cdot)$ is chosen as stated in (A1) i.e., $1/N \sum_{t=1}^{N} f_i(y(t)) = 0$. Otherwise, (A12) will be inconsistent with the fact that mean of $w(\cdot)$ is zero in eq. (1.1).

Finally consider the covariance matrix $\underset{\sim}{R}_i$ occurring in (A6) i.e., the probability density of the initial conditions $y(1),\ldots,y(m_i)$. The effect of this term on the final decision rule is not strong. Hence to simplify the computation, we make the assumption (A13).

(A13) $\quad \underset{\sim}{R}_i = \rho_{fi}\underset{\sim}{I}$

where ρ_{fi} = empirical variance of $f_i(y(t))$.

III. THE POSTERIOR PROBABILITY $P(C_i|\xi)$

Let $\xi = \{y(1),\ldots,y(N)\}$

$$P(C_i|\xi) = p(\xi|C_i)\ P(C_i)/p(\xi)$$

$$= \int_0^\infty d\rho \int_{\underset{\sim}{\theta}\in\Theta_i} d|\underset{\sim}{\theta}|\ p(\xi|\underset{\sim}{\theta},\rho,C_i)p(\underset{\sim}{\theta},\rho|C_i)P(C_i)/p(\xi) \qquad (3.1)$$

The expression for $p(\xi|\underset{\sim}{\theta},\rho,C_i)$ has been derived in section II. $p(\underset{\sim}{\theta},\rho|C_i)$ is available from assumptions (A8) and (A9). Hence the integration in (3.1) can be performed as indicated in the appendix I leading to the following theorem 1.

Theorem 1: Under the assumptions (A1)-(A13),the posterior probability $P(C_i|\xi)$ has the following form

$$P[C_i|\xi] = K \exp\ [0.5\ h_i(\xi)]$$

where

$$K = 1/\sum_{i=1}^{r} \exp[0.5\ h_i(\xi)]$$

$$h_i(\xi) = 2N\ \overline{\ln f_i'} - (N-m_i)\ \ln\ (\rho_i + (\rho_{fi}-\rho_i)/N)$$

$$+ 2\ \ln P(C_i) - n_i\ \ln N - m_i\ \ln \rho_{fi} + G_1(m_i) + O(1/N)$$

$$G_1(m_i) = m_i - (1/\rho_{fi}) \sum_{t=1}^{m_i} (f_i(t))^2$$

$$\rho_i = 1/(N-m_i) \sum_{t=m_i+1}^{N} (f_i(t)-\underset{\sim}{z}_i^T(t-1)\underset{\sim}{\theta}^*)^2$$

$$\underset{\sim}{\theta}^* = [(N-m_i+1)/(N-m_i)]\Big[\sum_{t=m_i+1}^{N} \underset{\sim}{z}_i(t-1)\underset{\sim}{z}_i^T(t-1)\Big]^{-1} \sum_{t=m_i+1}^{N} \underset{\sim}{z}_i(t-1)f_i(t)$$

$$\overline{\ln f_i'} = (1/N)\sum_{t=1}^{N} \ln f_i'(t),\ f_i'(t) = d\, f_i(y)/dy\Big|_{y=y(t)}$$

$$\rho_{fi} = (1/N)\sum_{t=1}^{N} (f_i(y(t)))^2 \triangleq \text{empirical variance of } \{f_i(y(t)),\ldots,f_i(y(N))\}$$

A proof of Theorem 1 is in appendix 1.

Comment 1: An expression for the $P(C_i|\xi(N))$ without the assumptions (A11)-(A13) regarding the parameters of prior distribution is given in lemma 1 in appendix 1. Obviously it is computationally more complicated.

Comment 2: $E[G_1(m_i)|C_i] = 0$

Variance $[G_1(m_i)|C_i] \approx 2m_i$

Hence $G_1(m_i)$ is of the order $0(\sqrt{m_i})$. While comparing 2 classes having different values of m_i, the term $G_1(m_i)$ does not make much difference in comparison with other terms and hence can be neglected when N is large.

Comment 3: The model $(f_i, g_i, \underset{\sim}{\theta}_i^*, \rho_i^*)$ is the best fitting model in the class C_i for the given data ξ. Alternatively, if ξ obeys some (unknown) model $(f_i, g_i, \underset{\sim}{\theta}, \rho)$ in C_i, then $\underset{\sim}{\theta}_i^*$ and ρ_i^* are the Bayesian estimates of $\underset{\sim}{\theta}$ and ρ.

We will discuss many other features of the posterior probability function $P(C_i|\xi)$ in section V.

IV. OPTIMAL DECISION RULE AND THEIR CONSISTENCY

1. Optimal Decision Rules:

Let $d(\xi(N))$ be a decision rule where d is a map:

$d: y^N \to \{C_1,\ldots,C_n\}$

Consider a loss function L which reflects the cost of wrong assignment of the observation set ξ to a class C_j using the rule.

$$L(C_i, d(\xi) = C_j) = 0 \text{ if } C_i = C_j$$

$$= w_{ij} > 0, \text{ if } C_i \neq C_j$$

Our intention is to choose the decision rule to minimize the average value of the loss function L, i.e., minimize J(d)

$$J(d) = E[L(C_i, d(\xi))]$$

$$= \sum_{i=1}^{r} P(C_i) \int L(C_i, d(\xi)) P(\xi|C_i) d|\xi|$$

or

$$J(d(\xi) = C_j) = \int \left(\sum_{i=1}^{r} w_{ij} P(C_i|\xi) \right) p(\xi)\, d|\xi|$$

The optimum decision rule is

$$d^*(\xi) = C_k = \underset{C_j \in \{C_1, \ldots, C_r\}}{\text{Argument Minimum}} \sum_{i=1}^{r} w_{ij} P(C_i|\xi) \qquad (4.1)$$

The loss function or, in particular, the weights w_{ij} can be chosen to reflect the particular needs of the problem. If we are interested in minimizing the probability of error in the assignment of class to ξ, the choice of w_{ij} is

$$w_{ij} = 0 \text{ if } i = j$$

$$= 1 \text{ if } i \neq j$$

In that case, the optimum decision rule is;

$$d^*(\xi) = C_k = \underset{C_j \in \{C_1, \ldots, C_r\}}{\text{Argument}} [\text{Maximum } P(C_j|\xi)] \qquad (4.2)$$

i.e. The decision rule assigns ξ to the class having the highest posterior probability.

$$\text{Probability of error of the optimal decision rule (4.2)} = [1 - \underset{C_j}{\text{Max}}\, P(C_j|\xi)] \qquad (4.3)$$

On the other hand, if our sole interest in class selection is to obtain a reliable estimate of spectral density, then we should choose w_{ij} as follows

$$w_{ij} = \int_{-\omega_N}^{\omega_N} d\omega (\ln S_i(\omega) - \ln S_j(\omega))^2$$

where $S_i(\omega)$ is the "best" estimate of the spectral density of the process based on ξ given that the process ξ belongs to class C_i. In that case, the optimum decision rule (4.1) simplifies

$$d^*(\xi) = C_j \quad \text{if } (\ln P(C_j|\xi)-H)^2 < (\ln P(C_i|\xi)-H)^2 \quad \forall\, i \neq j, \qquad (4.4)$$

with

$$H = \sum_{i=1}^{r} P(C_i|\xi)\ \ln P(C_i|\xi) \qquad (4.5)$$

The 2 illustrations should be sufficient to reveal the power of the decision rule to reflect the needs of the particular problem.

2. Consistency of the Decision Rule:

We will show that the optimum decision rule in (4.2) is asymptotically consistent. The consistency of other optimum decision rules such as (4.4) can be established in a similar manner. Without any loss of generality, let us consider the comparison of only 2 classes C_1 and C_2. If the process $y(\cdot)$ comes from some (unknown) model in class C_2, then we will show that $P(C_2|\xi(N))/P(C_1|\xi(N))$ tends to $+\infty$ as N tends to infinity showing that the decision rule correctly classifies the observation set. A precise expression for the $P(C_2|\xi(N))/P(C_1|\xi(N))$ is given in the following theorem 2.

We should emphasize that the asymptotic behavior of $P(C_2|\xi(N))/P(C_1|\xi(N))$ when C_2 is the correct class may be quite different from the asymptotic behavior of $P(C_1|\xi(N))/P(C_2|\xi(N))$ when C_1 is the correct class.

Theorem 2: Consider a pair of mutually exclusive classes C_1 and C_2, $C_i = \{f_i, g_i, m_i, n_i, \Theta_i\}$ under the assumptions (A1)-(A13). Assume that the given process $y(\cdot)$ obeys a model $\{f_2, g_2, \underset{\sim}{\theta}_2, \rho_2\} \in C_2$ where $\underset{\sim}{\theta}_2$ and ρ_2 are unknown, $\rho_2 > 0$.

Case (i) Let $f_1 = f_2$, (4.6)

$$n_1 > n_2 \tag{4.7}$$

For every $\underset{\sim}{\theta}' \in \Theta_2$, there exists a $\underset{\sim}{\theta} \in R^{n_1}$ so that (4.8) is satisfied.

$$g_1(t,\underset{\sim}{\theta}) = g_2(t,\underset{\sim}{\theta}') \tag{4.8}$$

Then

$$\lim_{N\to\infty} [P(C_2|\xi(N))/P(C_1|\xi(N))]^{1/\ln N} = \exp[(n_1-n_2)/2] \tag{4.9}$$

Case (ii) $f_1 = f_2$, but n_1, n_2 and g_1 and g_2 do not obey either (4.7) or (4.8). Then

$$\lim_{N\to\infty} [P(C_2|\xi(N))/P(C_1|\xi(N))]^{1/N} = k_2 > 1 \tag{4.10}$$

Case (iii) $f_1 \neq f_2$, then redefine the variable y so that $f_1(y) = y + k_1$, relabel $f_2(y)$ as $f(y)$. Assume $f(\cdot)$ obeys the following assumption (B1).

(B1) *If $f(y)$ is normal, then*

$$E[(y-\bar{y})^2] \geq E[(f(y)-\bar{f}(y))^2]/\exp[2E \ln f'(y)]$$

where $\bar{y}$ and $\bar{f}(y)$ stand for mathematical expectations of y and $f(y)$ and f' is the derivative of $f(y)$.

Then the limit given in (4.10) is also valid. Q.E.D.

The theorem is proved in appendix 2.

The assumption (B1) appears to be obeyed by most differentiable functions such as $f(y) = \ln y + k_3$, $f(y) = y^2 + k_3'$ etc. Still we have stated it as assumption since we do not have a proof of it.

Case (i) is valid for a pair of classes C_1 and C_2 where C_2 is obtained from C_1 after setting certain components of the parameter vector θ in C_1 to zero, and the true process obeys C_2. A common illustration is C_2 in a class of AR models of order n_2 and C_1 is a class of AR models of order n_1, $n_1 > n_2$, with the correct model belonging to C_2. In such a case, theorem 2 roughly states

$$P(C_2|\xi(N))/P(C_1|\xi(N)) \sim \exp(\frac{n_1-n_2}{2} \cdot \ln N) = N^{(n_1-n_2)/2}$$

or posterior error probability $\triangleq P(C_1|\xi(N)) = N^{-(n_1-n_2)/2}$ (4.11)

In all other cases, covered by Cases (ii) and (iii)

$P(C_2|\xi(N))/P(C_1|\xi(N)) \sim k_2^N$, $k_2 > 1$

or posterior error probability $= P(C_1|\xi(N)) \sim k_2^{-N}$ (4.12)

Eq. (4.11) clearly illustrates the empirically known fact that it is easier to distinguish 2 classes with entirely different structures, coming under Cases (ii) or (iii) than to discriminate between 2 classes with similar structures where the structure of the difference equation in C_2, the correct class can be obtained from that of C_1 by setting a few parameters in it to zero, i.e., it is difficult to distinguish the correct class C_2 from the higher order class C_1. Note that if C_1 and C_2 have similar structures, but $n_1 < n_2$, then we have an example of Case (i) and the error probability decays exponentially.

V. DISCUSSION AND COMPARISON

1. Main Characteristics of the Decision Rule:

We will highlight some of the important features of our decision rules. Most of these features are absent in the decision rules based on hypothesis testing and other methods which will be discussed subsequently.

(P1) The decision rule can compare simultaneously many classes obeying the conditions in Sections I and II.

(P2) The decision rule is transitive, i.e., Let $C_i \succ C_j$ denote that in comparing the classes C_i and C_j, the decision rule prefers C_i to C_j. Then $C_1 \succ C_2$ and $C_2 \succ C_3 \Rightarrow C_1 \succ C_3$ provided all the classes are equiprobable.

(P3) The decision rule is asymptotically consistent, i.e., while comparing 2 classes C_1 and C_2 based on the observation set $\xi(N)$, then $\lim_{N\to\infty} P(C_2|\xi(N))/P(C_1|\xi(N)) \to \infty$ if the observation set comes from some (unknown) member in C_2 and the error probability $P(C_1|\xi(N))$ decays at least as fast as $N^{-(n_1-n_2)/2}$ if $n_1 > n_2$ or k_2^{-N} where $k_2 > 1$.

(P4) The decision rule is optimal in the sense that it minimizes a suitable loss function which reflects the needs of the particular problem.

(P5) An explicit expression is available for the probability of error in the decision given by the rule.

(P6) The only arbitrary quantities appearing in the decision rule are the prior probabilities of the classes. There are no other arbitrary quantities like significance levels. The effect of the assumption about the prior probabilities is asymptotically negligible unlike the significance levels used in the hypothesis testing methods. Typically the prior probabilities of the classes can be assumed to be equal.

(P7) The roles played by m_i and n_i in the decision rules are quite different. The contribution of terms involving m_i to $\ln P(C_i|\xi)$ is $O(1)$ where as the contribution of terms involving n_i is $O(\ln N)$.

(P8) The posterior probability $P(C_i|\xi)$ involves a term $\exp[-n_i \ln N]$ even if all the r classes are a priori, equiprobable. When we are comparing 2 classes C_i, $i = 1,2$ such that they yield the same value of ρ_i and $\overline{\ln f_i'}$, $i = 1,2$, the posterior probability of the class having smaller n_i will tend to 1 as N tends to infinity. This is a quantitative proof of the "principle of parsimony" which states that if 2 models explain the same data (i.e., have same ρ), the model involving smaller number of estimated parameters has a higher plausibility of being the correct model of the process than the other model.

(P9) The decision rule for comparing classes is valid even if one of the members of a class yields a zero residual variance (i.e., $\rho_i = 0$) with the given data. This is possible because ρ always appears in the decision rule in the form $\ln[\rho + (\rho_f - \rho)/N]$. Hence even if ρ is zero or very small, $\ln[\rho + (\rho_f - \rho)/N]$ is still finite. This property is very useful in comparing stochastic models like autoregressive models involving a small number of parameters with polynomial

or fourier series models which involve a large number of parameters, but yield low residual variance.

2. Comparison with the Hypothesis Testing Approach:

This approach does not possess most of the properties (P2)-(P9). The hypothesis testing approach is usually designed to test whether certain components of vector $\underline{\theta}$ in the difference equation of form (1.1) are not significantly different from zero (hypothesis H_0 or null hypothesis) or the contrary (hypothesis H_1). Consequently the method can be used to compare only 2 classes at a time and these classes form a small subset of the classes mentioned in (P1). A typical application is when eq. (1.1) is a n_1 order autoregressive (AR) equation and the null hypothesis is that the given process $y(\cdot)$ obeys a n_2 order AR process, $n_2 < n_1$, i.e. the null hypothesis is that the coefficients of $y(t-n_2+1),\ldots,y(t-n_1)$ in (1.1) are all zero. Even here, we will show in example 4 that the decision rule is not always transitive (property P2) and the decision rule is not always asymptotically consistent (property P2). Hypothesis testing is routinely used in problems where the danger of rejecting H_0 when H_0 is true (type I error) is very much greater than accepting H_0 when H_0 is not true (type II error). The decision rule is designed to place an upperlimit on the probability of type I error, but no measure of the type II error is available (property P5). Thus the decision rule is inappropriate for problems when both types of errors are important. The decision rule involves an arbitrary parameter like significance level (i.e., probability of type I error allowed), (property P6). In general, the decision rule is not optimal in the sense that any specific criterion function is minimized (property P4). Further the hypothesis testing cannot handle a relatively simple comparison problem such as whether normal distribution or log normal distribution is appropriate for representing the given data or whether an autoregressive process or a polynomial fit (or

orthogonal function) is appropriate for representing the given data (property P9). Examples 1-4 bring out the limitations of the hypothesis testing approach.

3. Other Ad-Hoc Procedures:

There are many adhoc procedures for comparing several classes of model. The criterion MAIAC [3], is used for comparing classes of autoregressive models of different orders. For each class, we compute $a_i(\xi) = -N \ln \rho_i - 2n_i$, where

ρ_i = residual variance of the best fitting model in class C_i

n_i = number of parameters to be estimated.

Chosen class is C_i = Argument $\max_{C_i}[a_i(\xi)]$

Here there is no distinction between m_i and n_i. The term $-2n_i$ is inserted into the decision function using certain ideas of information theory, but there seems to be no particular reason for the factor 2. The decision rule is transitive. However, the decision rule has no optimality property. We have no idea of the probability of error of the decision rule. Most of the examples 1-4 are outside the scope of this rule. In particular, the rule cannot be used to compare classes in which one of the models such as a model with n orthogonal polynomials in t gives zero residual variance for the data since $a_i(\xi) = \infty$.

One can show that the use of decision rule is equivalent to the use of hypothesis testing procedure at an appropriate significance level [1].

Another adhoc approach is the maximum likelihood approach [2] in which we maximize the likelihood function in each class over the allowable set of parameters and choose that class which has the largest maximum likelihood value among them. As before the rule is not always asymptotically consistent and we have no idea of the probability of error given the rule. Further it is invalid for comparing 2 classes of AR models with different orders, since the maximum value of likelihood function associated with the class with larger

order is usually greater than that of the smaller order AR class. Similarly it cannot handle pairs of classes mentioned in (P9). The examples (1-4) bring out the relative merits of the various methods.

4. The Maximum Entropy Spectral Analysis [4,5,10]:

The original aim of the MESA approach is to obtain a reliable estimate of the spectral density of a process y from its N observations. Instead of assuming an arbitrary structure for the process, they found a structure for the process $y(\cdot)$ to maximize the entropy function under the following 2 restrictions

(i) $y(\cdot)$ is stationary and zero mean

(ii) All the correlations of the process up to mth order are known, i.e.,

ϕ_i, $i = 0,\ldots,m$ are known where $\phi_i = E[y(t)y(t-i)]$.

The result is an autoregressive structure for the process namely

$$y(t) = \sum_{t=1}^{m} a_j\, y(t-j) + w(t)$$

where w is a zero mean sequence $N(0,\rho)$ and $a_1,\ldots,a_m$ and ρ are determined from the known autocorrelations ϕ_i, $i = 0,\ldots,m$. The required spectral density is the $S(\omega)$

$$S(\omega) = \rho/||1 - a_1 e^{-i\omega} \ldots - a_{me}{}^{-im\omega}||^2$$

In practice, ϕ_i are not known and hence we replace them by the corresponding emperical correlation coefficients computed from the N observations $y(1),\ldots,y(N)$.

The key choice in the method is the integer m. The method does not suggest a method for the choice. The maximum value of m is N. Typically m can be 0.1N or 0.2N. The value of integer m is increased till the estimated spectral density shows sufficient resolution in the required frequency range.

On the other hand, if we use the approach of Section III for the selection of m, we have all the advantages of MESA method and the additional information

such as the posterior probability of C_i being correct class, an estimate of the variance of spectral density estimate etc. which are not available in MESA. In particular we can choose the decision rule to minimize errors in the estimation of spectral density as indicated in Section IV.

We will give 4 examples to illustrate the relative behavior of the various methods of comparison. In all the examples, the test observation set is $\xi = (y(1),\ldots,y(N))$. If C_i is the class, the corresponding residual variance is ρ_i and the corresponding signal variance is ρ_{fi}. $C_i \succ C_j$ means that the decision rule prefers the class C_i to C_j when the associated conditions are satisfied and the reverse if the conditions are not satisfied. The function $f_i(y)$ is chosen as mentioned in (A1).

Example 1: This example is used to illustrate the empirically observed fact that models which involve the estimation of a large number of parameters are usually inferior to appropriately chosen models involving a small number of parameters, even though the larger parameter model may result in zero residual variance. Specifically the 2 classes are

$$C_1 = \{f_1(y) = y+k,\ g_1,\ m_1 = 0,\ n_1 = N\}$$

$$C_2 = \{f_1,\ g_2,\ m_1 = 1,\ n_2 = 2\}$$

$$g_1(t,\underset{\sim}{\theta}) = \theta_1 + \sum_{k=1}^{N-1} \theta_{k+1}\phi_k(t-1)$$

$$g_2(t,y(t-1),\theta) = \theta_1 + \theta_2 y(t-1)$$

where $1,\phi_1(t)\ \phi_2(t)\ldots$ are a set of orthogonal functions orthogonal over $t = [1,N]$. They can be polynomials or sinusoids. C_2 is the class of first order AR models. Since our data set ξ has N observations, ρ_1, the residual variance of the best fitting model in class C_1 is zero. Let $\rho_{f1} = \rho_{f2} = \rho_y$ = empirical variance of $y(\cdot)$.

We can assume $\rho_2 >> \rho_y/N$. Decision rule (4.2) yields

$C_2 \succ C_1$, if $-(N-2) \ln \rho_2 - 2 \ln N - 2 \ln \rho_y > -N \ln(\rho_y/N) - N \ln N$

Retaining terms of order O(1) and higher, we get

$C_2 \succ C_1$ if $\rho_y/\rho_2 > \exp[(2 \ln N)/(N-2)]$

For instance, if N = 100

$C_2 \succ C_1$ if $\rho_2 < .91\ \rho_y$

i.e., as long as the first order AR model explains about 9 percent of the signal variance, the AR model class is superior to the class of models with N polynomials, inspite of the fact that residual variance $\rho_1 = 0$. The superiority of first or second order AR models to polynomial or orthogonal function models in modeling many (but not all) empirical series is well known to the workers dealing with empirical model building beginning with the work on sunspot series [9]. However the superiority of the AR model was demonstrated by an elaborate procedure like comparing correllograms and other validation methods [1]. In contrast the present theory offers a relatively simple quantitative explanation of this phonemenon.

Note that the conclusion is the same if we had compared C_1 with any other class C_3 of models involving 2 parameters, say, the class of models involving only 2 orthogonal functions, i.e., C_3 is the class of all straight line fits to the data in the plane $[t,y(t)]$ and C_1 is the class of polynomial fits. As before, we should prefer the straight line fit if it explains at least 9 percent of the signal variance in preference to the polynomial fit which yields zero residual variance.

Example 2: We compare 2 classes C_4 and C_5 having same n_i, but differ in f_i and possibly in g_i.

$C_4 = \{f_4, g_4, m, n\}$

$C_5 = \{f_5, g_5, m, n\}$

$f_4(y) = y + k_1$, $f_5(y) = \ln y + k_2$

Let ρ_{f4} = empirical variance of $y(1),\ldots,y(N)$

ρ_{f5} = empirical variance of $\ln y(1),\ldots,\ln y(N)$

$\overline{\ln y}$ = empirical mean of $\ln y(1),\ldots,\ln y(N)$

By decision rule (4.2)

$$C_5 \overset{i}{\gtrless} C_4 \text{ if } \ln(\rho_4/\rho_5) > 2\,\overline{\ln y}\left(\frac{N}{N-m}\right) + \frac{\ln(\rho_{f4}/\rho_{f5})}{N-m}$$

Note that the decision rule asymptotically does not depend on N explicitly. (It depends on N via ρ_i etc.).

As a particular illustration of this family of problems, let us determine whether a normal distribution or a log normal distribution is appropriate for representing the given observation set. Thus

$C_4 = \{f_4, g_4, m = 0, n = 1\}$

$C_5 = \{f_5, g_4, m = 0, n = 1\}$

$f_4(y) = y + k_1$, $f_5(y) = \ln y + k_2$

$g_4(t,\theta) = \theta_1$, $g_5(t,\theta) = \theta_1$,

Here $\rho_4 = \rho_{f4}$, $\rho_5 = \rho_{f5}$.

The decision rule (4.2) simplifies into

$$C_5 \overset{s}{\gtrless} C_4 \text{ if } \ln(\rho_{f4}/\rho_{f5}) > (2\,\overline{\ln y})/(1 + 1/N)$$

The simplicity of the decision rule should be compared with the corresponding complexity in using hypothesis testing methods. Using the hypothesis testing approach, we cannot directly compare the normal and log distribution fits to the given data. All we can do is compare whether the normal distribution $(\bar{y}, \rho_{f4})$ is

a good fit to the empirical distribution of the dataset at the 95 percent significance level. Alternatively we can test whether the log normal distribution $(\overline{\ln y}, \rho_{f5})$ is a good fit to the emperical distribution of $y(1),\ldots,y(N)$. It is not difficult to construct examples in which we can find both the normal and log normal fits are significant at the 95 percent significance level.

Example 3: (Autoregressive processes) We will compare 2 classes of autoregressive models of orders n_6 and n_7 respectively.

$C_i = \{f, g_i, n_i, n_i\}$, $i = 6,7$, $n_7 > n_6$

$f(y) = y$, ρ_f = empirical variance of y

$\nabla_\theta g_i(t,\underset{\sim}{\theta}) = (y(t-1),\ldots,y(t-n_i))$

Let $N >> n_6$, $N >> n_7$ and $\rho_i >> \rho_f/N$, $i = 6,7$

Decision rule (4.2) yields:

$C_7 \rangle C_6$ if

$$-N \ln \rho_7 - n_7 \ln(\rho_f/\rho_7) - n_7 \ln N > -N \ln \rho_6 - n_6 \ln(\rho_f/\rho_7) - n_6 \ln N$$

i.e., $C_7 \rangle C_6$, if $$\ln(\rho_6/\rho_7) \rangle \frac{(n_7-n_6)[\ln N + \ln(\rho_f/\rho_6)]}{N - n_7} \qquad (5.1)$$

If ρ_6 and ρ_7 are not too far from each other

$$\ln(\rho_6/\rho_7) \simeq (\rho_6-\rho_7)/\rho_7$$

i.e., we should increase the order of AR model from n_6 to n_7 only if the fractional decrease in variance is greater than the quantity on the RHS of (5.1). If $n_7 = n_6+1$, $N = 200$, $(\rho_f/\rho_6) = 2$ then RHS of (5.1) = .03155 i.e., we should add another AR term only if the fractional decrease in variance is at least 3.15 percent.

Suppose we want to use the hypothesis testing approach. Then the decision rule has the following form:

$$C_7 \rangle C_6 \text{ if } \frac{\rho_6-\rho_7}{\rho_7} \rangle \frac{(n_7-n_6)k(n_6-n_7,N)}{N} \qquad (5.2)$$

where k is a threshold depending on n_6-n_7,N and the significance level. It is determined by the fact that if C_6 is true then

$$N(C_6-C_7)/\rho_7(n_7-n_6) \sim F(n_6-n_6,N)$$

For N > 100, the threshold depends only on (n_7-n_6).

Thus the principle difference between rules (5.1) and (5.2) is the absence of the factor ln N and the presence of the nonlinear threshold k in (5.2).

Example 4: Let C_i denote class of AR models of order i in which $n_i = i$. Our intention is to show that the decision rule given by hypothesis testing (i.e. rule (5.2) is intransitive (i.e. it does not obey P2). Let ρ_i denote the residual variance of class C_i. Let N = 100. Suppose we choose 95 percent significance level.

$$k(1, 100) = 3.84 = k_1$$

$$k(4, 100) = 2.38 = k_4$$

We will presently show that the nonlinear dependence of $k(n_1,n_2)$ on n_1 is the cause of the intransitivity.

If we compare C_i and C_{i+1}, (5.2) yields

$$\begin{array}{l} \dfrac{N(\rho_i-\rho_{i+1})}{\rho_{i+1}} \leq k_1 \Rightarrow \text{choose } C_i \\ \dfrac{N(\rho_i-\rho_{i+1})}{\rho_{i+1}} > k_1 \Rightarrow \text{choose } C_{i+1} \end{array} \qquad (5.3)$$

Suppose we compare C_i and C_{i+4}

$$\begin{array}{l} \dfrac{N(\rho_i-\rho_{i+4})}{4\rho_{i+4}} \leq k_4 \Rightarrow \text{choose } C_i \\ \dfrac{N(\rho_i-\rho_{i+4})}{4\rho_{i+4}} > k_4 \Rightarrow \text{choose } C_{i+4} \end{array} \qquad (5.4)$$

Suppose the numerical values of ρ_i, i = 1,2,...,5 obey the following relations (5.5) and (5.6) which are not inconsistent

$$\frac{4k_4}{N} + 1 < \frac{\rho_1}{\rho_5} < \left(1 + \frac{k_1}{N}\right)^4 \qquad (5.5)$$

$$\rho_i/\rho_{i+1} \leq \frac{k_1}{N} + 1, \quad i = 1,\ldots,4 \qquad (5.6)$$

The decision rule (5.4) and the left hand side of inequality (5.5) imply (5.7)

$$C_5 \succ C_1 \tag{5.7}$$

Decision rule (5.3) and eq. (5.6) imply (5.8)

$$C_i \succ C_{i+1} \ , \ i = 1,\ldots,4 \tag{5.8}$$

Repeated use of (5.8) implies (5.9)

$$C_1 \succ C_5 \tag{5.9}$$

Eq. (5.9) and (5.7) mutually contradict each other showing the intransitivity of the decision rule (5.2).

Note, however, that the decision rule (5.1) is not intransitive since it does not involve any arbitrary threshold.

VI. CONCLUSION

We have developed a method of comparing different classes of dynamical models using Bayesian theory. The method can handle a wide variety of classes and is much superior to the traditional methods of comparison like hypotheses testing. The method clearly shows the limitations of models such as polynomial fits which using a large number of parameters can render the residual variance zero. It clearly shows that such models have no explanatory power.

REFERENCES:

[1] R. L. Kashyap and A. R. Rao, Dynamic Stochastic Models from Empirical Data, Academic Press, 1976.

[2] E. L. Lehmann, Testing Statistical Hypothesis , Wiley, New York, 1959.

[3] H. Skaike, "A New Look at Statistical Model Identification", IEEE Trans. on Automatic Control, Vol. AC-19, Dec. 1974, p. 716-722.

[4] D. E. Smylie, G. K. C. Clarke and T. J. Ulrych, "Analysis of Irregularities in the Earth's Rotation" in "Methods in Computational Physics, vol. 13, Geophysics, (ed.) Alder et al.

[5] J. P. Burg, "The Relationship Between Maximum Entropy Spectra and Maximum Likelihood Spectra", Geophysics, vol. 37, p. 375-376.

[6] M. S. Bartlett, An Introduction to Stochastic Processes, Cambridge Univ. Press, London, 1966.

[7] J. D. Markel, "Digital Inverse Filtering: A New Tool for Formant Trajectory Estimation", IEEE Trans. Audio Electroacoustics, AU-20, 1972, p.129-137.

[8] Kashyap, R. L., "Probability and Uncertainty", IEEE Trans. Information Theory, 1971, pp. 641-650.

[9] G. U. Yule, "On the Method of Investigating Periodicities in Disturbed Series with Special Reference to Wolfer's Sunspot Numbers", Phil. Trans. A226, 1927, p. 267.

[10] N. Andersen, "On the Calculation of Filter Coefficients for Maximum Entropy Spectral Analysis", Geophysics, vol. 49, no. 1, Feb. 1974, pp. 69-72.

Appendix 1

We will establish theorem 1 in 3 steps. We will first state lemma 1 which gives an expression for $P(C_i|\xi)$ using only the assumptions (A1)-(A10). We shall state lemma 2. Using lemma 1 and lemma 2, and the additional assumptions (A11)-(A13), we will prove theorem 1. Next, we will establish lemmas 3, 1, 2 successively, lemma 3 being required in the proof of lemma 1.

Lemma 1:

Under the assumptions (A1)-(A10), the posterior probability of $P(C_i|\xi)$ has the following expression

$$P(C_i|\xi) = k \exp[0.5h_i(\xi) + O(1/N)], \quad i = 1,\ldots,r \tag{1}$$

where

$$h_i(\xi) = 2n\lambda_i - (N-m_i)\ln\beta_i - n_i \ln N + \ln(\det \bar{S}_i/\det \bar{S}_{0i}) - \ln \det R_i + 2\ln p(C_i) + b_i, \tag{2}$$

$$b_i = m_i - (f_i(1),\ldots,f_i(m_i))R_i^{-1}(f_i(1),\ldots,f_i(m_i))^T \tag{3}$$

$$\lambda_i = (1/N) \sum_{t=1}^{N} d \ln f_i(y)/dy\Big|_{y=y(t)} \tag{4}$$

$$\beta_i = \hat{\rho}_i + (1/(N-m_i))(\tilde{\theta})^T S_{0i}^{-1}\tilde{\theta} \tag{5}$$

$$\tilde{\theta} = \theta^* - \theta_{0i} = S_i \sum_{t=m_i+1}^{N} z_i(t-1)(f_i(t) - z_i^T(t-1)\theta_{0i}) \tag{6}$$

$$S_i = [S_{0i}^{-1} + \sum_{t=m_i+1}^{N} z_i(t-1)\, z_i^T(t-1)]^{-1} \tag{7}$$

$$\bar{S}_i = S_i(N-m_i) \tag{8}$$

$$\hat{\rho}_i = (1/(N-m_i)) \sum_{t=m_i+1}^{N} (f_i(t) - z_i^T(t-1)\theta_i^*)^2 \tag{9}$$

$$z_i(t-1) = \nabla_\theta g(t,y(t-1),\ldots,y(t-m_i),\theta) \tag{10}$$

Lemma 2:

Under the assumptions (A11) and (A12) β_i in (5) will simplify as follows

$$\beta_i = \rho_i + (\rho_{fi}-\rho_i)/(N-m_i) \tag{11}$$

where

$$\rho_{fi} = 1/N \sum_{t=1}^{N} (f_i(t))^2 \tag{12}$$

Proof of Theorem 1:

Let us first use the expression for S_{0i} in (A11) which simplifies S_i in (7) as follows

$$S_i = S_{0i}/(N-m_i+1) \tag{13}$$

we can use (13) to simplify the expression $\ln(\det \bar{S}_i/\det S_{0i})$ occurring in (2). Using (8) and (13), we get

$$\ln \det \bar{S}_i - \ln \det S_{0i} = \ln \det[S_{0i} \frac{(N-m_i)}{N-m_i+1}] - \ln \det S_{0i}$$

$$= n \ln(N-m_i/N-m_i+1)$$

$$\simeq -n/(N-m_i) = 0(1/N) \tag{14}$$

Next, let us use assumption (A13). Then b_i in (3) simplifies into (15)

$$b_i = m_i - \sum_{t=1}^{m_i} (f_i(t))^2/\rho_{fi} \tag{15}$$

clearly $E[b_i|C_i] = 0$.

$$\text{Variance } [b_i|C_i] = 2m_i$$

Moreover, (A13) implies $\ln \det R_i = m_i \ln \rho_{fi}$ (16)

Substituting (11), (14)-(16) in (2), yields theorem 1. Q.E.D.

We need the following lemma 3 to prove lemma 1. To simplify the notation, we will drop the subscript i hereafter, i.e., denote m_i, θ_i^*, $\hat{\rho}_i$ etc. by m, θ^*, $\hat{\rho}$, etc.

Lemma 3:

$$\sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta)^2 + (\theta-\theta_0)^T S_0^{-1}(\theta-\theta_0)$$

$$= (\theta-\theta^*)^T S^{-1}(\theta-\theta^*) + (N-m)\beta \qquad (17)$$

Proof of Lemma 3: Let $\tilde{\theta} = \theta^*-\theta_0$

$$\sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta)^2 + (\theta-\theta_0)^T S_0^{-1}(\theta-\theta_0)$$

$$= \sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta^* + z^T(t-1)(\theta^*-\theta))^2$$

$$+ (\theta-\theta^* + \tilde{\theta})^T S_0^{-1}(\theta-\theta^* + \tilde{\theta})$$

Expanding the squares and rearranging terms

$$\text{LHS of (17)} = [(\theta^*-\theta)^T(\sum_{t=m+1}^{N} z(t-1)z^T(t-1) + S_0^{-1})(\theta^*-\theta)]$$

$$+ \sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta^*)^2 + (\tilde{\theta})S_0^{-1}\tilde{\theta}]$$

$$+ [2(\theta^*-\theta)^T\{\sum_{t=m+1}^{N} z(t-1)(f(t) - z^T(t-1)\theta^*) - S_0^{-1}\tilde{\theta}\}], \qquad (18)$$

Coefficient of $(\theta^*-\theta)^T$ in (18)

$$= \sum_{t=m+1}^{N} z(t-1)f(t) - \sum_{t=m+1}^{N} z(t-1)z(t-1)\theta^* - S_0^{-1}\tilde{\theta}$$

$$= \sum_{t=m+1}^{N} z(t-1)(f(t) - z^T(t-1)\theta_0) - [\sum_{t=m+1}^{N} z(t-1)z^T(t-1) + S_0^{-1}]\tilde{\theta}$$

$$= 0, \text{ by definition of } \tilde{\theta} \text{ in (6) and (7)} \qquad (19)$$

Substitute (19) in (18) and rewrite R.H.S. of (18) using definition of β in (5) and (9). We get

LHS of (17) = RHS of (17).

Proof of lemma 1:

Recall $\xi_1 = \{y(1),\ldots,y(m_i)\}$, $\xi_2 = \{y(m_i+1),\ldots,y(N)\}$.

$$\text{Let } k = \prod_{t=1}^{N} (df(y)/dy)\Big|_{y=y(t)} \tag{20}$$

We will first compute $p(\xi_2|\xi_1,\rho,C)$

$$p(\xi_2|\xi_1,\rho,C) = \int d\underset{\sim}{\theta} p(\xi_2|\xi_1,\theta,\rho,C)p(\theta|\rho,C,\xi_1)$$

$$= \int d\underset{\sim}{\theta} p(\xi_2|\xi_1,\theta,\rho,C)p(\theta|\rho,C), \text{ by (A7)} \tag{21}$$

$$= k \int d|\underset{\sim}{\theta}| \frac{1}{(2\pi\rho)^{(N-m)/2}} \exp[-1/2\rho) \sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta)^2]$$

$$\frac{1}{(2\pi)^{n/2}} \frac{1}{(\text{Det } S_0\rho)^{1/2}} \exp[-\frac{1}{2\rho} (\theta-\theta_0)^T S_0^{-1}(\theta-\theta_0)],$$

using (2.4) and (A9).

$$= k \int d|\theta| \frac{1}{(2\pi\rho)^{(N-m+n)/2}} \frac{1}{(\text{Det } S_0)^{1/2}} \exp[-(1/2\rho)$$

$$\{\sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta)^2 + (\theta-\theta_0)^T S_0^{-1}(\theta-\theta_0)\}], \text{ by rearranging terms}$$

$$= k \int d|\theta| \frac{1}{(2\pi\rho)^{(N-m+n)/2}} \cdot \frac{1}{(\text{Det } S_0)^{1/2}} \exp[-(1/2\rho)(\theta-\theta^*)^T$$

$$S^{-1}(\theta-\theta^*)] \exp[-(1/2\rho)(N-m)\beta] \tag{22}$$

using lemma 3.

$$= k \frac{1}{(2\pi\rho)^{(N-m)/2}} \frac{(\text{Det } S)^{1/2}}{(\text{Det } S_0)^{1/2}} \exp[-(1/2\rho)(N-m)\beta],$$

Now we will integrate over ρ after multiplying by $p(\rho|C)$ given by (A8)

$$p(\xi_2|\xi_1,C) = \int d\rho\ p(\rho|C)p(\xi_2|\xi_1,\rho,C)$$

$$= \int d\rho \frac{\alpha}{\rho} \frac{k}{(2\pi\rho)^{(N-m)/2}} \frac{(\text{Det } S)^{1/2}}{(\text{Det } S_0)^{1/2}} \exp[-(1/2\rho)(N-m)\beta]$$

$$= 2k\alpha \int_0^\infty dx \exp[-x^2(N-m)\beta/2)]x^{N-m-1}(\text{Det } S/\text{Det } S_0)^{1/2}/(2\pi)^{(N-m)/2},$$

using the transformation $\rho = 1/x^2$,

$$= \alpha k[(N-m)\beta/2]^{-(N-m)/2}\ \Gamma[(N-m)/2][\text{Det } S/\text{Det } S_0]^{1/2}/(2\pi)^{(N-m)/2}$$

Simplify the above expression using the expression for k in (20) and the standard formula for $\Gamma(x)$ namely:

$$\ln\Gamma(x) = x \ln x - x + 1/2 \ln(2\pi/x),$$

$$2 \ln p(\xi_2|\xi_1,C) = 2 \ln \alpha + 2 \sum_{t=m+1}^{N} \ln f'(t) - (N-m) \ln \beta - (N-m)$$

$$+ \ln[4\pi/(N-m)] + \ln(\det S/\det S_0) - (N-m)\ln 2\pi, \quad (23)$$

$\underset{\sim}{S}$ is of the order $O(1/N)$. Hence we can define $\bar{S} = S(N-m)$, so that $\bar{S}$ is $O(1)$.

$$\det S = (\det \bar{S})(1/N-m)^n \quad (24)$$

By assumption (A6)

$$2 \ln p(\xi_1|C) = - m \ln 2\pi - \ln \det R + 2 \sum_{t=1}^{m} \ln f'(t) - ||(f(1),\ldots,f(m)||_R^2 - 1 \quad (25)$$

$$2 \ln p(\xi|C) = 2 \ln p(\xi_2|\xi_1,C) + 2 \ln p(\xi_1|C) \quad (26)$$

$$= 2 \sum_{t=1}^{N} \ln f'(t) - (N-m) \ln \beta - (n+1) \ln (N-m)$$

$$+ \ln(\det \bar{S}/\det S_0) - \ln \det R + m - ||f(1),\ldots,f(m)||_R^2 - 1$$

$$- N \ln 2\pi - N + 2 \ln\alpha + \ln(4\pi), \quad \text{using (23) and (25)}, \quad (27)$$

$$\ln p(C|\xi) = \ln p(\xi|C) + \ln p(C) - \ln p(\xi)$$

$$= 0.5h(\xi) + \text{terms independent of any class} + O(1/N) \quad (28)$$

where

$$h(\xi) = 2 \sum_{t=1}^{N} \ln f'(t) - (N-m) \ln\beta - n \ln N + \ln(\det \bar{S}/\det S_0)$$

$$- \ln \det R + 2 \ln p(C) + (m - ||f(1),\ldots,f(m)||_R^2 - 1$$

Proof of lemma 2:

Recall $\rho_f = (1/N-m) \sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta_0)^2$

Recall that

$$\rho = (1/N-m) \sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta^*)^2$$

$$= (1/N-m) \sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta_0 - z^T(t-1)\tilde{\theta})^2, \text{ by definition of } \tilde{\theta} \text{ in (6)}$$

$$= (1/N-m) \sum_{t=m+1}^{N} (f(t) - z^T(t-1)\theta_0)^2 + (\tilde{\theta})^T (1/N-m) \sum_{t=m+1}^{N} z(t-1)z^T(t-1)\tilde{\theta}$$

$$- 2\tilde{\theta}^T \sum_{t=m+1}^{N} z(t-1)(f(t) - z^T(t-1)\theta_0)/N-m$$

Substitute for the coefficient of $\tilde{\theta}$ in the last term using (6), replace the first term by ρ_f and rearrange the terms.

$$\rho = \rho_f + \tilde{\theta}^T (1/N-m \sum_{t=m+1}^{N} z(t-1)z^T(t-1) - 2S^{-1}/N-m)\tilde{\theta}$$

$$= \rho_f + (\tilde{\theta})^T S_0^{-1} \tilde{\theta} (1 - 2(N-m+1)/N-m)$$

or $\tilde{\theta}^T S_0^{-1} \tilde{\theta} = (\rho_f - \rho)/(1 + 2/N-m)$

$$\approx (\rho_f - \rho)(1 - 2/N-m)$$

By definition of β in (5),

$$\beta = \rho + (1/N-m)(\tilde{\theta})^T S_0^{-1} \tilde{\theta}$$

$$= \rho + (1/N-m)(\rho_f - \rho) + 0(1/N)$$

Appendix 2

We will outline the proof of theorem 2:

Let $(f_i, g_i, \theta_i^*(N), \rho_i^*(N)) \in C_i$, $i = 1,2$, be the best fitting models for the given data in the 2 classes C_1 and C_2, θ_i^*, ρ_i^*, etc., being defined in Section III. We will assume that $\theta_i^*(N)$ and $\rho_i^*(N)$ tends to θ_i and ρ_i as N tends to infinity with probability one. In view of the assumptions (A1)-(A7), the models in each class are identifiable given the class. Hence $(f_2, g_2, \theta_2, \rho_2)$ the asymptotically recovered model is the exact representation of $y(\cdot)$ stated in the theorem. Note that by definition, $y(\cdot)$ does not obey the model $(f_1, g_1, \theta_1, \rho_1)$. Rather, this model is the best fitting model in C_1 for the semiinfinite data set $\{y(1),y(2),\ldots\}$.

Recall that

$$\ln[p(C_2|\xi(N)/p(C_1|\xi(N))] = 0.5[h_2(\xi(N)) - h_1(\xi(N))] \qquad (1)$$

Cases (i) and (ii)

Without any loss of generality, we can set $f_1(y) = f_2(y) = y + k_1$

$\hat{y}_1(t|t-1)$ = one step ahead predictor of $y(t)$ based on $y(t-j)$, $j \geq 1$ suggested by the model $(f_1, g_1, \underset{\sim}{\theta}_1, \rho_1)$

$$= g_1(t,\theta) - k_1 \qquad (2)$$

$$E[y(t)|y(t-j),\ j \geq 1] = g_2(t,\underset{\sim}{\theta}_2) - k_1$$

By definition of expectation and normality of y

$$E[y(t) - \hat{y}_1(t|t-1))^2] \geq E[(y(t) - \hat{y}_2(t|t-1))^2]$$

i.e., $\rho_1 \geq \rho_2$

At this point, we will discuss the cases (i) and (ii) separately.

Case (i): The structure of g_1 and g_2 mentioned in this case imply (3)

$$\rho_1 = \rho_2 \tag{3}$$

$$\lim_{N\to\infty} \frac{1}{\ln N} \ln[P(C_2|\xi(N))/P(C_1|\xi(N))]$$

$$= \lim_{N\to\infty} 0.5 \frac{(h_2(\xi(N)) - h_1(\xi(N))}{\ln N}$$

$$= \lim_{N\to\infty} \frac{0.5}{\ln N} \{-N \ln \hat{\rho}_2 + N \ln \hat{\rho}_1 + (n_1-n_2)\ln N + O(1)\}$$

$$= \lim_{N} \frac{0.5}{\ln N} \{-N \ln \rho_2 + N \ln \rho_1 + (n_1-n_2)\ln N + O(1)\}, \text{ since } \hat{\rho}_1 \to \rho_1 \text{ and}$$

$\hat{\rho}_2 \to \rho_2$ w.p.1.

$= (n_1-n_2)/2$, by (3)

or $(P(C_2|\xi(N))/P(C_1|\xi(N))^{1/\ln N} \to \exp(\frac{n_1-n_2}{2})$,

In Case (ii):

$\rho_1 > \rho_2$

or $\rho_1 = \rho_2 \exp[k_3]$, $k_3 > 0$

$$\lim_{N\to\infty} \frac{1}{N} \ln[P(C_2|\xi(N))/P(C_1|\xi(N))] = \lim_{N\to\infty} (\ln \hat{\rho}_1 - \ln \hat{\rho}_2) = k_2 > 0,$$

or $(P(C_2|\xi(N))/P(C_1|\xi(N))^{1/N} \to \exp[k_3] \triangleq k_2 > 1$

Case (iii):

Without any loss of generality we can set $f_1(y) = y + k_1$

Define $y_1(t|t-1)$ as in (2).

By definition

$$\rho_1 \triangleq E[(y(t) - y_1(t|t-1))^2|y(t-j), j \geq 1]$$

$$\geq E[\{y(t) - E(y(t))|y(t-j), j \geq 1)\}^2|y(t-j), j \geq 1] \tag{4}$$

Let $\bar{f}_2(t|t-1) = E[f(y(t))|y(t-j), j \geq 1]$

$$= g_2(t, \underline{\theta}_2)$$

Since the conditional distribution of $f(y(t))$ given $y(t-j), j \geq 1$ is normal, we can use condition (B1) and rewrite it as follows:

$$E[(y(t) - E(y(t)|y(t-j), j \geq 1))^2 | y(t-j), j \geq 1]$$

$$\geq E[(f(y(t)) - \bar{f}_2(t|t-1))^2 | y(t-j), j \geq 1] / \exp[2E(\ln f'(y(t)|y(t-j), j \geq 1)] \quad (5)$$

Using (4)

$$\ln \rho_1 > \ln E[\{y(t) - E(y(t)|y(t-j), j \geq 1)\}^2 | y(t-j), j \geq 1$$

$$\geq \ln \rho_2 \, 2E(\ln f'(y(t))|y(t-j), j \geq 1)]$$

using (5) and definition of ρ_2

$$\geq \ln \rho_2 \, 2E(\ln f'(y(t)) \quad (6)$$

$$\lim_{N\to\infty} \frac{1}{N} \ln[P(C_2|\xi(N))/P(C_1|\xi(N))]$$

$$= \lim_{N\to\infty} [0.5 \, [h_2(\xi(N)) - h_1(\xi(N))]/N]$$

$$= \lim_{N\to\infty} [2(1/N) \sum_{t=1}^{N} \ln f_2'(y(t)) - \ln \rho_2 + \ln \rho_1 + O(\log N/N)]$$

$$\geq k_3 > 0, \text{ by (6)}$$

or

$$((P(C_2|\xi(N))/P(C_1|\xi(N)^{1/N} \sim \exp[k_3], \text{ for large } N$$